# A Bayesian Approach to Concept Drift

Stephen H. Bach, Marcus A. Maloof

## Introduction

- *Concept drift* occurs in a sequential classification task when the target concept $f: X \to Y$, i.e., the true mapping from attribute values to class labels, changes.
- Many ensemble methods for drift train members on (possibly overlapping) blocks of consecutive examples.
- Such methods address directly the uncertainty about the existence and location of drift.
- We place a probability distribution over the location of the most-recent drift point and use it to weight the influence of ensemble members when making predictions.

## Bayesian Model Comparison

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

- Adams and MacKay (Technical Report, University of Cambridge, 2007) used Bayesian model comparison to reason about the location of the most-recent change point in sequential observations assumed to be generated by a non-stationary distribution.
- Placed a prior over $l_t$, the location of the most-recent change point at time step $t$.

$$p(l_t|l_{t-1}) = \begin{cases} \lambda^{-1} & \text{if } l_t = 0; \\ 1 - \lambda^{-1} & \text{if } l_t = l_{t-1} + 1; \\ 0 & \text{otherwise.} \end{cases}$$

- Marginalized $p(l_t, l_{t-1}|D_{1:t})$ to obtain $p(l_t|D_{1:t})$.
- Looked for changes in the joint distribution over all observed features $D$.

## Comparing Conditional Distributions

- Our goal is to model the conditional distribution as accurately as possible, so we model the location of the most recent *drift point* in that conditional distribution.
- We use the method of Adams and MacKay as a starting point.
- Marginalize $l_{t-1}$ from

$$p(l_t, l_{t-1}|Y_{1:t}, X_{1:t}) = \frac{p(Y_t|l_t, Y_{1:t-1}, X_{1:t})p(l_t|l_{t-1})p(l_{t-1}|Y_{1:t-1}, X_{1:t-1})}{p(Y_t|Y_{1:t-1}, X_{1:t})}$$

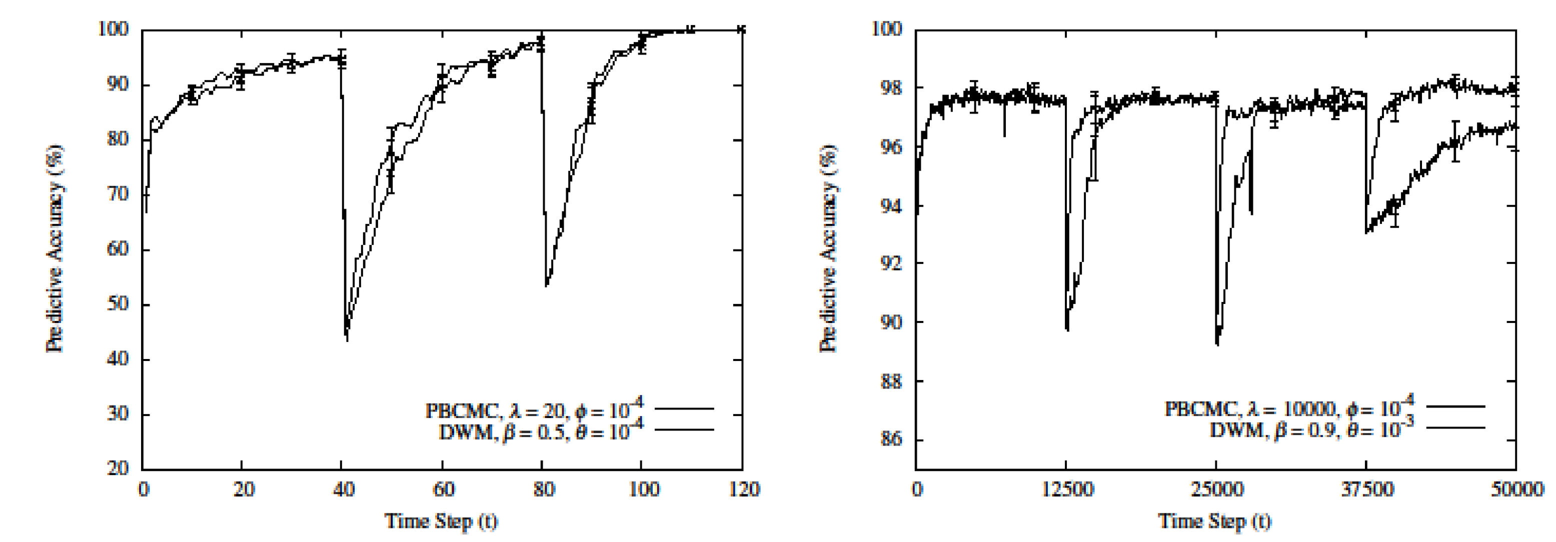
- Prune possible locations of drift when their probabilities fall below a threshold $\phi < p(l_t = 0|l_{t-1})$.

## Empirical Evaluation

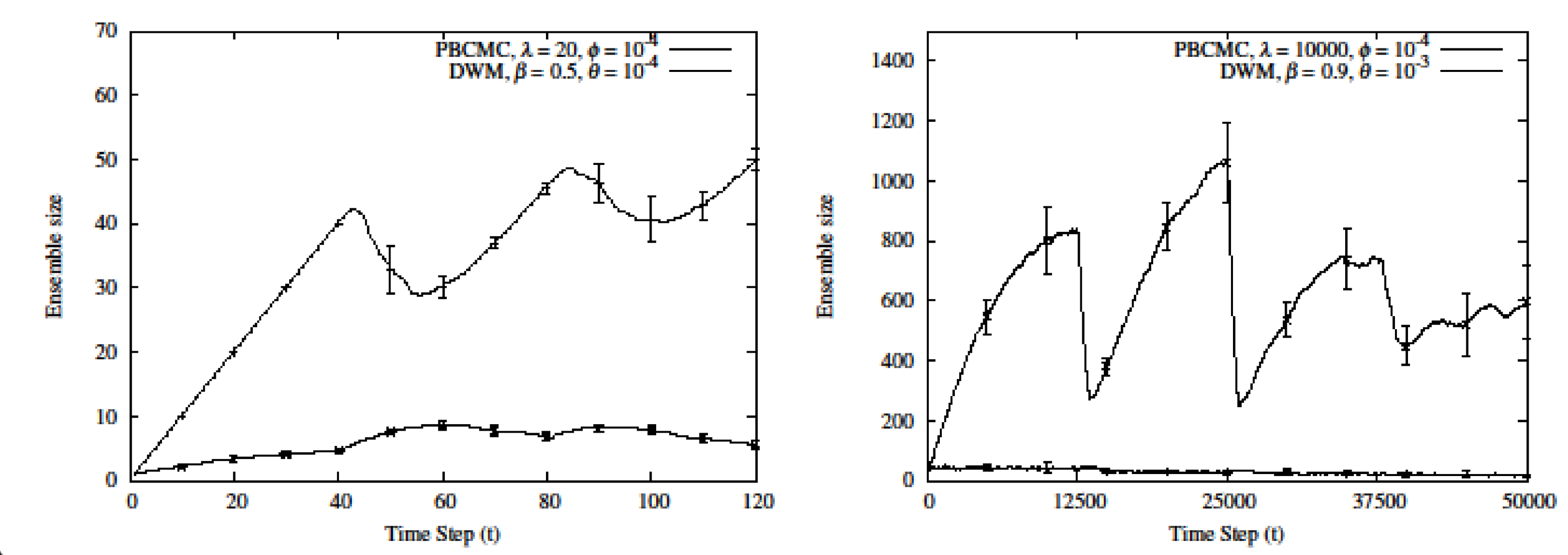
- Classifiers:
  - PBCMC – Our model.
  - BCMC – Our model without pruning.
  - BMC – Model of Adams and MacKay. Compares accuracies of joint distributions.
  - Dynamic Weighted Majority - A leading ensemble method for concept drift.
  - Bayesian Naïve Bayes – Base learner for all ensembles. Places Dirichlet priors over discrete distributions and Normal-Gamma priors over continuous distributions. Evaluated as base line.
- We tested the learners on two synthetic problems, the STAGGER and SEA concepts, and two real-world problems, the CAP and electricity prediction data sets.
- PBCMC and BCMC outperformed BMC, often dramatically.
- PBCMC and DWM each offered different benefits on different problems.

## Selected Results on STAGGER and SEA Concepts

- Accuracy:




- Ensemble size:




## Conclusions

- Looking for drift points in the conditional distributions, rather than change points in the joint distributions, led to much greater accuracy in our experiments.
- PBCMC and DWM each had different advantages when either more reactivity or more stability was desirable.

This work was supported by the Georgetown University Undergraduate Research Opportunities Program.

Department of Computer Science, Georgetown University http://cs.georgetown.edu